Towards Scalable Assessment of Performance-Based Skills: Generalizing a Detector of Systematic Science Inquiry to a Simulation with a Complex Structure

Michael A. Sao Pedro[1,2], Janice D. Gobert [1,2], and Cameron G. Betts [2]

[1] Learning Sciences & Technologies Program, Worcester Polytechnic Institute, Worcester, MA
Apprendis LLC, Stow, MA USA
{mikesp, jgobert}@wpi.edu
[2] Apprendis LLC, Stow, MA USA
cam@apprendis.com

Abstract. There are well-acknowledged challenges to scaling computerized performance-based assessments. One such challenge is reliably and validly identifying ill-defined skills. We describe an approach that leverages a data mining framework to build and validate a detector that evaluates an ill-defined inquiry process skill, designing controlled experiments. The detector was originally built and validated for use with physical science simulations that have a simpler, linear causal structure. In this paper, we show that the detector can be used to identify demonstration of skill within a life science simulation on Ecosystems that has a complex underlying causal structure. The detector is evaluated in three ways: 1) identifying skill demonstration for a new student cohort, 2) handling the variability in how students conduct experiments, and 3) using it to determine when students are off-track before they finish collecting data.

Keywords: science simulations, science inquiry, inquiry assessment, performance assessment, behavior detector, reliability, educational data mining

1 Introduction

Performance-based assessment tasks, complex tasks that require students to create work artifacts and/or follow processes, are being seen as alternatives to multiple-choice questions because the latter have been criticized as not capturing authentic and relevant "21st century skills" such as critical and creative thinking (e.g. [1]), and scientific inquiry (e.g. [2]). When implemented using computerized simulations [3], games [1] and virtual worlds [2], they have the potential to be scaled because they can be deployed consistently, can automatically evaluate students' work products and processes they follow to create those work products [1], [2], [3], [4], and by virtue of automatic assessment, can provide real-time feedback to students and educators [1], [3]. However, an assessment challenge arises when skills are ill-defined (cf. [1]), meaning that there are many correct or incorrect ways for students to demonstrate skills [5]. How can assessment designers guarantee that the evaluation rules or models

they author [1] to identify demonstration of skill within a given task are consistently and accurately doing so? Furthermore, how can they guarantee models will work across different contexts (tasks)?

In this paper, we explore the challenge of creating reliable, scalable evaluation of an ill-defined scientific inquiry process skills in the context of Inq-ITS [3], a simulation-based intelligent tutoring system that also acts as a performance assessment of students' inquiry skills. We determine whether an evaluation model (detector) of an inquiry process skill already shown to generalize for physical science simulations with simple, linear causal structures [6], [7], [8], [9] can also identify the skill in a Life Sciences simulation on Ecosystems that has a complex causal structure (cf. [10]).

2 Prior Work: Validating a Designing Controlled Experiments Detector for Inq-ITS Physical Science Activities

Inq-ITS [3] is a web-based virtual lab environment in which students conduct inquiry with interactive simulations and inquiry support tools. The simulations were designed to tap content areas aligned to middle school Physical, Life, and Earth Science described in Massachusetts' curricular frameworks. Each Inq-ITS activity provides students a driving question, and requires them to investigate that question using the simulation and tools (see Figure 1 for an example Ecosystems activity) in a semi-structured inquiry. More specifically, students attempt to form a testable hypothesis using a pulldown menu-based sentence builder, collect data by changing the simulation's variables and running trials (Figure 1), analyze their data using pulldown menus to construct a claim and by selecting trials as evidence, and communicate findings in an open text field (see [3]). A key aspect of the system is that activities are performance assessments of inquiry skill, because skills are inferred from the inquiry processes they follow and the work products they create with the support tools.

The process skill of focus in this paper is *designing controlled experiments* when collecting data with the simulation. Students design controlled experiments when they generate trials that make it possible to infer how changeable factors (e.g. seaweed, shrimp, small fish, and large fish within an Ecosystem) affect outcomes (e.g., the overall balance of the ecosystem) [6]. This skill relates to application of the Control of Variables Strategy (CVS; cf. [11]), but unlike CVS, it takes into consideration *all* the experimental design setups run with the simulation, not just isolated, sequential pairs of trials [6], [3]. The challenge in assessing this skill is that it is ill-defined; students' data collection patterns can vary widely and there are many ways to successfully demonstrate (or not demonstrate) this process skill [12]. The added difficulty of conducting inquiry in a complex system whose variables interact in nonlinear ways (as opposed to simpler linear systems in which variables have more straightforward dependencies [13]) also contributes to the multitude of ways in which students collect data. This in turn also affects the complexity of assessing this skill.

To address this assessment difficulty, we developed and validated a data-mined detector to determine whether students designed controlled experiments within Inq-ITS physical science activities [6], [7], [8], [9]. We chose a data mining approach to over-

come limitations of other models that could under- or over-estimate students' mastery of this skill (e.g. [14]), and to enable easier validation of how well it would perform by testing it against data not used to build it, thereby addressing issues of reliability and scalability (see [12], [9] for a discussion). Data mining was applied to build models that could replicate human judgment of whether or not students designed controlled experiments. Training and testing labels were generated using text replay tagging of students' log files [15], [6], a process in which human coders tag segments of logfiles (clips) with behaviors or skills. This detector was originally built for a physical science topic on Phase Change as a J48 decision tree. In subsequent work, the decision tree was further improved by choosing features that increased the theoretical construct validity of the detector, and by iterative refinement of the decision tree to find an optimal feature set [7], [9]. Examples of chosen features included the number of data trials collected, how many times the simulation variables were changed, various counts of controlled trials in which only one variable was changed, and various counts for repeated trials with the exact same simulation setup. The detector uses cutoffs of feature values to predict if a student designs controlled experiments.

Overall, we have strong evidence for using this detector to evaluate the designing controlled experiments skill for physical science inquiry activities at scale. For example, as well as being able to predict skill demonstration on held-out test data for Phase Change (the same student sample and simulation from which it was constructed [7]), the models also generalized to predict the same skill within two other physical science topics on energy during free fall [8] and density [9]. The generalization test to the Energy activities also addressed how well the model could handle both new students, and the variability in how they collect data and demonstrate skill [8]. The detector was also validated for a second purpose, determining if a student was off-track when collecting data [7]. In follow-on work, the detector was deployed in Inq-ITS to drive proactive interventions, *before they finished collecting data* in the Phase Change simulation [16], [12]. Thus, the detector could both assess the skill when students finish collecting data, and to drive interventions.

The present study extends this prior work to determine if this detector built and validated for physical science simulations can evaluate the skill and drive interventions for a more complex Life Science simulation on Ecosystems. We adapt our former analytical techniques [6], [7], [8], [9] to address this question.

3 Inq-ITS EcoLife Ecosystems Activities

The EcoLife simulation assesses students' inquiry skills and hones their knowledge of ecosystems. It addresses the two strands of the Massachusetts Curricular Frameworks: 1) the ways in which organisms interact and have different functions within an ecosystem to enable survival, and 2) the roles and relationships among producers, consumers, and decomposers in the process of energy transfer in a food web. The EcoLife simulation (Figure 1) consists of an ocean ecosystem containing big fish, small fish, shrimp, and seaweed. Two inquiry scenarios were developed for this simulation. In the first, students are explicitly told to stabilize the ecosystem. In the second, stu-

dents are to stabilize the shrimp population (or alternatively, ensure that the shrimp population is at its highest). Students then address the questions by engaging in the inquiry process described earlier.

There are key differences between our physical science simulations and the Ecosystems simulation that can make assessing the designing controlled experiments skill more difficult. For example, unlike the physical science simulations that have discrete choices for variable values [3], in Ecosystems students add and remove organisms with varying numbers. The Ecosystems simulation model is also complex causal system whose multiple variables are interconnected in a non-linear fashion [13], [10], unlike the physical science simulations which have simple linear dependencies [3]. This added complexity increases the hypothesis search space [17], and makes understanding the effects of the independent variables on dependent variable(s) more challenging. As such, the simple control for variables strategy (cf. [11]) may not be applied in a straightforward manner for this task.

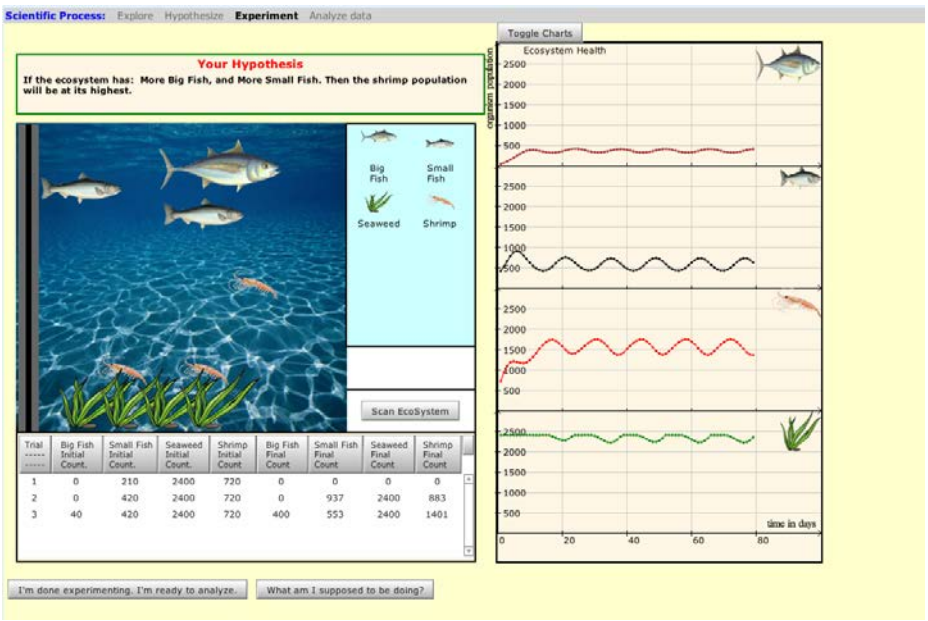

Fig. 1. EcoLife experiment stage. Here, students add and remove organisms, and scan the ecosystem to determine how the population changes over time.

4 Dataset: Distilling Clips from Ecosystems Activities

We collected interaction data from 101 students from a Central Massachusetts middle school who engaged in inquiry with the Ecosystems activities. Then, text replay tagging of log files (clips) [15] was again used to generate a test set for evaluating the applicability of the detector to Ecosystems. A clip contains all actions associated with

formulating hypotheses (hypothesize phase actions) and all actions associated with designing and running experiments (experiment phase) [6].

One human coder (the third author) tagged all the clips distilled from the Ecosystems logfiles. A second coder who originally tagged clips in physical science also tagged the first 50 clips to test for agreement. Aside from training the first coder, determining inter-rater reliability was particularly important because, in addition to its complexity, the Ecosystems environment has a substantially different UI and interaction pattern than the previous physical science simulations [3]. Agreement for the 50 clips tagged by both coders was high overall, $\kappa = .71$, on par with our prior work coding for this skill [6]. In total, 226 clips were tagged, and of those, 52.2% were tagged as the student having demonstrated skill at designing controlled experiments.

5 Results: Generalizability of the Detector to Ecosystems

The overarching goal of this paper is to determine how well the designing controlled experiments detector built and validated for physical science simulations with a simpler, linear causal model, generalizes to predict skill demonstration in a second topic, Ecosystems with a more complex simulation. This goal is important to ensure the model can correctly identify skill in multiple simulation contexts, students and students' experimentation patterns. To do so, three questions are addressed: First, acknowledging that there might be individual differences in how students conduct inquiry in general, can the detector be applied to new students who used the Ecosystems simulation [8]? Second, can the detector handle the variability in how students collect data in Ecosystems [8]? Finally, can the detector be used to determine when scaffolding could be applied when a student is "off-track" [7]?

Commensurate with our prior work on testing the goodness of detectors [6], [7], [8], [9], the degree to which the detector agrees with human judgment (the clip labels described previously) is summarized using two metrics, A' computed as the Wilcoxon statistic [18] and Cohen's Kappa. Briefly, A' is the probability that the detector can distinguish a clip where skill is demonstrated from a clip where skill is not demonstrated, given one clip of each kind. The chance value of A' is .50. Cohen's Kappa (κ) estimates whether the detector is better than chance ($\kappa = 0.0$) at agreeing with the human coder's judgment. A' and Kappa were chosen because, unlike accuracy, they attempt to compensate for successful classifications occurring by chance (cf. [19]). A' can be more sensitive to uncertainty in classification than Kappa, because Kappa looks only at the final label, whereas A' looks at the classifier's degree of confidence.

5.1 Can the detector be applied to new students in Ecosystems?

The following analysis benchmarks how well the detector handles new students in the new science domain with a more complex simulation [8]. As mentioned earlier, this cohort of students came from a different school than those from which the original detector was built. As shown in Table 1, the detector's performance was quite high and indicate that the detector can be used to evaluate new students' performance in

the Ecosystems activities [8]. It could distinguish when a student designed controlled experiments in Ecosystems from when they did not A' = 75% of time. The detector's overall agreement with human judgment of whether a student designed controlled experiments was also quite high, κ = .61. This performance is on par with previous metrics computed at the student-level across three physical science topics, A' ranging from .82 to .94 and κ ranging from .45 to .65 across studies [7], [8], [9].

Table 1. Confusion matrix and performance metrics computed when applying the designing controlled experiments detector to the Ecosystems clips.

	True N	True Y
Pred N	91	27
Pred Y	17	91

Pc = .84, Rc = .77
K = .61, A'=.75

* Pc = precision; Rc = recall

Table 2. Performance metrics for the designing controlled experiments detector disaggregated by number of trials in students' experimentation.

Runs	# Clips	A'	K	Pc	Rc
[2,3]	40	.90	.76	.83	.83
[4,5]	39	.64	.44	.78	.67
[6-10]	38	.53	.07	.82	.60
>10	65	.66	.20	.88	.89

* Pc = precision; Rc = recall

5.2 Can the detector handle the variability in how students collect data?

Though the previous results are highly encouraging, they only reveal one aspect of generalizability. We found in prior work that by sampling data according to the variability in students' experimentation patterns, specifically how many trials they collected, we could reveal weaknesses in the detector [8]. We follow a similar process here to characterize how well the detector handles the experimentation variability within Ecosystems. Unlike [8] in which clips were sampled to balance exact counts of trials collected by students (e.g. clips where students collected exactly 4 trials, clips with exactly 5 trials, etc.), here clips were binned into different groups of variability. As an example, one bin contained 40 clips where students collected exactly 2 or 3 trials (Table 2). This deviation was performed because there was greater variability in the number of trials run by students in Ecosystems than in Physical Science. In addition, the number of clips for any specific number of runs was not large enough to generate valid performance metrics. Bins were chosen to both balance the number of clips per bin and to ensure each had enough set of clips for generating metrics.

As shown in Table 2, the detector handled the variability in students' experimentation reasonably well. Performance was high for clips with 2 or 3 simulation runs (A'

= .90, κ = .76) and clips with 4 or 5 runs (A' = .64, κ = .44). The detector did, however, struggle on predicting clips with 6 to 10 runs as indicated by A' = .53 and κ = .07 values close to chance. It also did not perform as well for clips with more than 10 runs, A' = .66 and κ = .20, albeit better than chance.

5.3 Can the detector identify when students are "off-track" when designing controlled experiments so that scaffolds can be effectively applied?

As mentioned, it is also of interest to determine if the detector can be used to identify when students are off-track by not designing controlled experiments. This is important so that a timely intervention can be given *before* they finish collecting data to prevent floundering [16]. We can determined this by measuring how well the detector can identify skill using less data than was used by the human coder to identify skill [7]. More specifically, we can use a subset of a student's interaction data up to and including the *nth* time the student ran the simulation to predict if a student ultimately did/did not design a controlled experiment. The grain size of "simulation run" was chosen because an intervention given at this point may prevent students from floundering and collecting more confounded data [3], [16].

Like [7], detector performance was measured using data up to a given number of simulation runs. Since there was more variation in how many times the simulation was run in Ecosystem and its increased complexity, detector performance was measured by varying the number of simulation runs from 1 to 10. Again, A' and κ were computed for each simulation run. As shown in Figure 2, the detector can predict if a student is "off-track" when collecting data in Ecosystem in as few as 3 simulation runs, indicated by A' and κ values well above chance, replicating earlier findings [7]. We note the detector performs at chance level for exactly one simulation run because the designing controlled experiments skill can be only identified after the student has collected two or more trials with the simulation (cf. [11]). We also note, however, that as the number of runs exceeds 6, the detector has difficulty distinguishing positive from negative examples. This is indicated by A' values ranging from .58 to .66. The detector, though, still agrees with human judgment fairly well, κ = .41 to .52. The implications of this finding are discussed in the next section.

6 Discussion and Conclusions

Performance-based assessments (e.g. [1], [2], [3]) present added assessment challenges when the underlying skills they tap are ill-defined (cf. [1]). The main challenge is that such skills may be demonstrated in many correct or incorrect ways by the student (e.g. [5]) which calls to question the reliability and applicability of the underlying assessment models aimed at identifying such skills. Towards the goal of providing reliable, scalable performance-based assessment of inquiry, we determined if a data-mined detector for designing controlled experiments [6], originally built for Physical Sciences simulations [7], [8], [4] that have simpler, linear dependencies between simulation variables, could be applied to the same skill in Ecosystems, a more complex

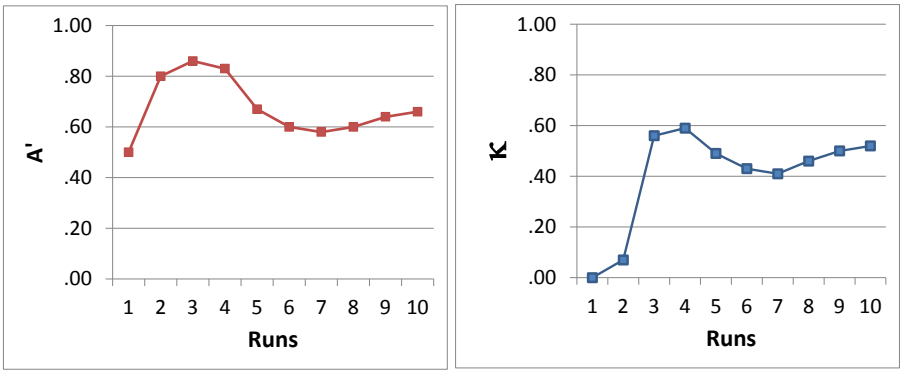

Fig. 2. Designing controlled experiments performance (A' and κ) predicting skill demonstration using data up to and including the *nth* simulation run, n = [1,10]. As shown, the detector can be applied in as few as three simulation runs. However, as the number of runs exceeds 6, the detector has difficulty time distinguishing positive from negative examples (indicated by A' closer to chance = .5) even though it still agrees well with human labels (κ >= .40).

simulation. In brief, we addressed if the detector could: 1) handle student-level validation, 2) assess the multi-faceted ways in which students' conduct inquiry in a complex system, and 3) predict when scaffolding in this domain is needed, a question of importance since the system aims to provide feedback to students as they experiment to prevent them from floundering [3], [16].

The results indicated that the detector had broad generalizability (cf. [20]) given that it could reliably assess the skill within Ecosystems and given its prior success at doing so for physical science simulations [7], [8], [4]. Its performance on the Ecosystems data was akin to that of the physical science simulations [7], [8], [4] under student-level validation. When assessing variability of how students experimented, the detector could identify skill demonstration well when students ran between 2 and 5 trials, but performance dropped when students collected more data than 5 trials. Finally, we found evidence that the detector could detect if a student was "off track" in as few as three simulation runs, commensurate with prior findings within a physical science simulation [7], but also had lower performance as the number of runs increased above 5. One possible way to overcome this limitation as the number of runs increases is to reset the 'window' of students' experimentation patterns after they receive scaffolding, i.e., after a student receives scaffolding, the system could treat the student as if they had not conducted any actions with the simulation. Then, after three more data collections, the system could again determine if the student is still off-track.

This work makes two contributions towards performance-based assessment and generalizability of EDM detectors. First, this study complements prior work on building generalizable detectors of affect (e.g. [21]) and other undesirable behaviors within ITS's (e.g. [20], [22]) with its focus on skill assessment. The power of using the EDM approach to build models that identify skill demonstration is in the ability to *learn* evaluation rules (cf. [1]) from student data, and the ability to quantify how reliable the model is at identifying skills for new students and within different tasks (e.g. physical science vs. life science) by testing detector performance with new student data. Sec-

ond, as in [7], [8], [9], this study employs additional validation techniques in addition to student-level generalizability tests (e.g. [21], [22]) to determine the extent to which the detector can be used to evaluate skill and drive scaffolding in the more complex domain of Ecosystems. While student-level validation is important, other aspects specific to assessment such as handling variability in how students engage in performance-based tasks and specific to formative assessment such as students get timely feedback so they do not flounder [3] are also necessary if such models are to generalize to multiple situations. Overall, these results are promising towards realizing scalable assessment and real-time formative feedback of inquiry skill development across science topics. In particular, our computer-based approach complements other assessments of deep science knowledge (e.g. [23]) by focusing on inquiry skills. In addition, since our assessments are performance-based, they may help overcome the limitations associated with assessing inquiry via traditional methods [2].

The generalizability and reusability of the detector has been hypothesized to be due to judicious feature engineering [7]. As such, including other types of features may improve prediction and generalizability. For example, [8] suggests that using ratio-based features instead of a raw counts for features may improve generalizability. For future work, issues such as improved feature engineering will be explored to ensure this detector can work for new students, handle the variability in students experiment, and ensure that scaffolding will be applied at an appropriate time across all Inq-ITS activities for physical, life, and earth science.

Acknowledgements

This research is funded by the National Science Foundation (NSF-DRL#0733286, NSF-DRL#1008649, and NSF-DGE#0742503) and the U.S. Department of Education (R305A090170 and R305A120778). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies.

References

1. Shute, V.: Stealth Assessment in Computer-Based Games to Support Learning. In : Computer Games and Instruction. Information Age Publishing, Charlotte, NC (2011) 503-523
2. Clarke-Midura, J., Dede, C., Norton, J.: The Road Ahead for State Assessments., Policy Analysis for California Education and Rennie Center for Educational Research & Policy, Cambridge, MA (2011)
3. Gobert, J., Sao Pedro, M., Baker, R., Toto, E., Montalvo, O.: Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. Journal of Educational Data Mining 4(1), 111-143 (2012)
4. Rupp, A. A., Gushta, M., Mislevy, R. J., Shaffer, D. W.: Evidence-centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments. The Journal of Technology, Learning, and Assessment 8(4), 1-45 (2010)
5. Shute, V., Glaser, R., Raghavan, K.: Inference and Discovery in an Exploratory Laboratory. In : Learning and Individual Differences: Advances in Theory and Research. W.H. Freeman, New York, NY (1989) 279-326

6. Sao Pedro, M. A., Baker, R. S. J. d., Gobert, J. D., Montalvo, O., Nakama, A.: Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. User Modeling and User-Adapted Interaction 23, 1-39 (2013)
7. Sao Pedro, M., Baker, R., Gobert, J.: Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. In : Proc. of the 20th Conf. on User Modeling, Adaptation, and Personalization, Montreal, QC, Canada, pp.249-260 (2012)
8. Sao Pedro, M. A., Baker, R. S. J. d., Gobert, J. D.: What Different Kinds of Stratification Can Reveal about the Generalizability of Data-Mined Skill Assessment Models. In Proc. of the 3rd Conference on Learning Analytics and Knowledge, Leuven, Belgium (2013)
9. Gobert, J., Sao Pedro, M., Raziuddin, J., Baker, R.: From Log Files to Assessment Metrics for Science Inquiry using Educational Data Mining. Journal of the Learning Sciences 22(4), 521-563 (2013)
10. Greiff, S., Wustenberg, S., Funke, J.: Dynamic Problem Solving: A New Measurement Perspective. Applied Psychological Measurement 36, 189-213 (2012)
11. Chen, Z., Klahr, D.: All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. Child Development 70(5), 1098-1120 (1999)
12. Sao Pedro, M.: Real-time Assessment, Prediction, and Scaffolding of Middle School Students' Data Collection Skills within Physical Science Simulations. Ph.D. Dissertation etd-042513-062949, Worcester Polytechnic Institution, Worcester, MA (2013)
13. Yoon, S.: An Evolutionary Approach to Harnessing Complex Systems Thinking in the Science and Technology Classroom. Int'l Journal of Science Education 30(1), 1-32 (2008)
14. McElhaney, K., Linn, M.: Helping Students Make Controlled Experiments More Informative. In : Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences, Chicago, IL, pp.786-793 (2010)
15. Baker, R. S. J. d., Corbett, A. T., Wagner, A. Z.: Human Classification of Low-Fidelity Replays of Student Actions. In : Proceedings of the Educational Data Mining Workshop held at the 8th International Conference on Intelligent Tutoring Systems, ITS 2006, Jhongli, Taiwan, pp.29-36 (2006)
16. Sao Pedro, M., Baker, R., Gobert, J.: Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition. In : Proc. of the 6th International Conference on Educational Data Mining, Memphis, TN, pp.185-192 (2013)
17. van Joolingen, W. R., de Jong, T.: An Extended Dual Search Space Model of Scientific Discovery Learning. Instructional Science 25, 307-346 (1997)
18. Hanley, J. A., McNeil, B. J.: The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143, 29-36 (1982)
19. Ben-David, A.: About the Relationship between ROC Curves and Cohen's Kappa. Engineering Applications of Artificial Intelligence 21, 874-882 (2008)
20. Baker, R. S. J. d., Corbett, A. T., Roll, I., Koedinger, K. R.: Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction 18(3), 287-314 (2008)
21. Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C.: Population Validity for Educational Data Mining Models: A Case Study in Affect Detection. To appear in the British Journal of Educational Technology (accepted)
22. San Pedro, M. O. Z., Baker, R., Rodrigo, M. M.: Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics. In : Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, NZ, pp.304-311 (2011)
23. Liu, O., Lee, H., Linn, M. C.: Multifaceted Assessment of Inquiry-Based Science Learning., 69-86 (2010)